SUPPLEMENTARY MATERIALS

Calibrating variant-scoring methods for clinical decision making

Silvia Benevenuta[1], Emidio Capriotti [2,] and Piero Fariselli [1,*]*
[1]Department of Medical Sciences, University of Torino, Via Santena, 19, 10126 Torino, Italy.
[2]BioFolD Unit, Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Via Selmi 3, 40126 Bologna, Italy

* Correspondence should be addressed to emidio.capriotti@unibo.it and piero.fariselli@unito.it

Supplementary Methods

1. The benchmark dataset

The dataset of single nucleotide variants (SNVs) used to compare the different methods was extracted from ClinVar (Landrum, *et al*., 2016) (http://www.ncbi.nlm.nih.gov/clinvar/) annotated with hg19. The SNVs were filtered by selecting only the ones with either Pathogenic or Benign annotation that are present in two recent versions of ClinVar (December 2018 and June 2019) and have not been included in PhD-SNP^{g} training sets. After the filtering procedure, we ended up with a set of 2,607 Pathogenic and 1,033 Benign SNVs. For testing purposes, the final set is composed of all the Benign SNVs and an equal number of randomly selected Pathogenic variants. The composition of the dataset is reported in Table S1.

2. Calibration techniques and plot

In this study, we considered two different calibration technique based either on isotonic or sigmoid mapping functions (Niculescu-Mizil and Caruana, 2005). The latter mapping procedure is more effective when the distortion in the predicted probabilities is sigmoid-shaped. The isotonic calibration is extremely powerful, but it is prone to overfitting if the training dataset is not sufficiently large. To plot the calibration curves of each method, we partitioned our dataset with either a uniform or quantile strategy. The quantile-based approach was shown to be more robust than the uniform one. More information about the calibration procedures and plots are described in Supplementary Materials.

3. Overview of the prediction methods

In this study, we considered the following six predictors:

- CADD (Kircher, *et al*., 2014) is an integrative annotation built on more than 60 genomic features, and it can score human single nucleotide variants and short insertion and

deletions anywhere in the reference assembly. For improved interpretability, the scores are transformed into a PHRED-like rank score. CADD training data consist of 16 million observed variants and 49 million simulated variants. It can annotate both coding and non-coding variants.

- DANN (Quang, *et al*., 2015) uses the same feature set and training data of CADD to train a deep neural network (DNN). Its output is a probability and predicts both coding and non-coding variants.
- DeepSea (Zhou and Troyanskaya, 2015) has been developed to predict only the effect of non-coding variants. It directly learns a regulatory sequence-code from large-scale chromatin-profiling data, enabling prediction of chromatin effects of sequence alterations with single-nucleotide sensitivity. The predictor was trained on a diverse compendium of genome-wide chromatin profiles from the Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics projects. It returns the functional significance score for each variant.
- Eigen (Ionita-Laza, *et al*., 2016) is an unsupervised approach and is not based on any labelled training data. It uses a variety of functional annotations in both coding and non-coding regions (such as made available by the ENCODE and Roadmap Epigenomics projects) and combines them into one single measure of functional importance. Eigen does not provide the scores for the variants on the X and Y chromosomes.
- FATHMM-MKL (Shihab, *et al.*, 2015) integrates functional annotations from ENCODE with nucleotide-based sequence conservation measures. It is trained on two distinct datasets: the Human Gene Mutation Database (Stenson, et al., 2003) and the 1000 Genomes Project (1000 Genomes Project Consortium, *et al.*, 2012). It can be used both on coding and non-coding variants and it provides a probability.
- PhD-SNP[g] (Capriotti and Fariselli, 2017) is a classifier based on a limited amount of sequence information and on conservation scores. It is trained on the ClinVar dataset and can classify both coding and non-coding variants. The output of this method is a probabilistic score.

4. How to plot the calibration curve

The function to plot the calibration curve can be found in the python library *scikit-learn* (Pedregosa, F. *et al.*, 2011) (see: https://bit.ly/31TwQbr to understand its parameters).

It follows this procedure:

- Take the binary label and its predicted probability given by the model;
- Sort the data using the probability given;
- Create bins in two possible ways:
  - **i) Uniform**: divide the interval [0,1] into subsets of fixed length, and consider each subset as a bin (note that these bins might have a very different number of elements);
  - **ii) Quantile**: divide the range of probabilities [0,1] into quantiles and consider each subset as a bin;
- Calculate the fraction of positive labels in each bin and the average of the probabilities and plot them, using the average probabilities on the x-axis and the fraction on the y-axis.

5. *Calibration of the classifiers*

In Fig. S1**,** S2 and S7 we report the calibration curves of the methods using the **uniform** strategy (as one can see, they are quite unstable); the calibration curves with the **quantile** strategy are in Figs. S3-S6. In Fig. S1 and S3 we compared the calibration of the classifiers that have a probability as output, without modifying the scores in any way (we assumed that probabilities should have intrinsic meaning); in Fig. S2 and S4 we compared the calibrated scores of Eigen and CADD after an isotonic calibration; while in Fig. S6 and S7 we performed a sigmoid calibration.
We plotted the calibration curves separately on coding, non-coding and all the variants to see how they behave differently depending on the strength of the conservation signal (coding variants have a stronger conservation signal). Before plotting the calibration curves in Fig. S5, we applied the isotonic calibration to the raw scores of Eigen and CADD training the calibration with 10-fold cross-validation, which means that we divided the dataset into 10 subsets, fit the calibration on 9 of these subsets and transformed the scores of the 10th subset. This operation was repeated 10 times so that every variant had its score transformed without overfitting and the transformed scores were assembled all together to plot the calibration curve.
We also investigated the effect of the isotonic regression (cv10) on all the classifiers: the calibrations improved greatly both on coding and non-coding variants (the effects are reported in Fig. S14-S19). Whenever possible (datasets bigger than 1000 elements), we suggest using this type of calibration.

6. *Techniques for calibrating the probabilities*

Prediction probabilities are calibrated using the following techniques:

- **Sigmoid/ Platt's:** The mapping function is a sigmoid with parameters to be determined**.** The function is defined as: $P_i = \frac{1}{1+e^{Af_i+B}}$, where $f_i = f(x_i)$ is the output of the model given x_i, $P_i = \mathbb{P}(x_i)$ is the probability of the data point x_i after the calibration and A and B are parameters to be learned. Platt Scaling is most effective when the distortion in the predicted probabilities is sigmoid-shaped;

- **Isotonic Regression:** This method is more general since the only restriction is that the mapping function is isotonic (monotonically increasing). That is: $y_i = m(f_i) + \epsilon_i$, given the predictions f_i and the true targets y_i, where m is an isotonic function. The problem is finding the isotonic function such that $m = argmin_z \sum(y_i - z(f_i))^2$. This calibration is a more powerful calibration method that can correct any monotonic distortion but is more prone to overfitting when data is scarce (and thus performs worse than Platt Scaling in that situation).

7. *Brier score*

The Brier score is defined as:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2,$$

where i is an item in a set of N predictions, f_i is the predicted probability of the element i and o_i is the actual outcome. Therefore, the lower the Brier score is for a set of predictions, the better the predictions are calibrated. It takes on a value between zero and one since it is the same range in which the probabilities and the labels vary. This score is appropriate for binary, categorical outcomes, but can not be used for ordinal variables or for cases with three or more classes.

8. Calculation of the receiver operating characteristic curve

In all the performance measures - assuming that positives indicate Pathogenic and negatives indicate Benign - TP (true positives) are correctly predicted Pathogenic Single Nucleotide Variants (SNVs), TN (true negatives) are correctly predicted Benign variants, FP (false positives) are Benign SNVs annotated as Pathogenic, and FN (false negatives) are Pathogenic variants predicted to be Benign.
The area under the receiver operating characteristic (ROC) curve (AUC), is obtained by plotting the True Positive Rate (TPR) as a function of the False Positive Rate (FPR) at different probability thresholds of annotating a variant as Pathogenic or Benign.

$$TPR = TP/(TP + FN)$$
$$FPR = FP/(FP + TN)$$

9. Simulated data

We wanted to further explore the effect and the power that the calibration had on the scores and how and when they could actually be calibrated using the sigmoid function and the isotonic calibration. We asked ourselves if there were distributions of the scores that could not be influenced by the calibration and how effective would the standard calibration be. Thus, we simulated different types of distributions, calibrated the scores on the training set and then tested the different calibrations on a separated set. In the Figures S8, S10, S11, it is possible to see how the calibration process would work in the case of a perfect classifier (Fig. S8), when handling a very polarized distribution with quite a few errors near the endpoints (Fig. S10) and when treating an almost-random classifier (Fig. S11). In Figure S9 we show how a model could be almost perfectly calibrated without necessarily being a perfect classifier (therefore, we want to emphasize the importance of controlling the AUC before going on with further inquiries).

10. Performance of the methods

In Table S2 we report the AUCs values for each method on coding and non-coding variants. Note that DeepSea was not constructed to work on coding variants. CADD is the best performing classifier on both coding and non-coding variants, with PhD-SNP[g] as a close second.

11. Calibration scores

Not all classifiers we studied provide well-calibrated probabilities. Here, we wanted to show the distribution of the scores/probabilities of the six algorithms to have a better understanding of their ill-calibration. As shown in the Fig. S12 and S13, the scores of the benign variants tend to be widespread from zero to one, while the scores of the pathogenic tend to be consistently close to one and this reflects badly on the calibrations. The developers of the algorithms should check the distributions of the outcomes while working on them, to obtain better-calibrated models.

12. Calibration of Eigen and CADD with a sigmoid function

We also tried to calibrate the scores of Eigen and CADD using a sigmoid function

$$f(x) = \frac{1}{1+e^{-x+B}}$$

The parameter B was chosen using the threshold the maximized the Matthews Correlation Coefficient (MCC), a measure of the quality of binary classifications. We transformed and calibrated CADD's scores with the coefficient $B = 2.5$ that was the threshold for the best classification on both coding and non-coding variants. We calibrated the scores from Eigen separately on coding and non-coding SNVs since it provides two different sets of scores. We use $B = 0.05$ for the coding variants and $B = 1.63$ for the non-coding variants. After this transformation, CADD showed a good level of calibration, while Eigen still performed very poorly. The calibrations of Eigen and CADD with this procedure are shown in Fig.S6 and S7 (uniform and quantile).

Supplementary Tables

Table S1. Composition of the dataset used for testing the methods

Variant Type	Pathogenic	Benign
Non-Coding	228	476
Synonymous	9	364
Missense	796	193
Total	1,033	1,033

Table S2. Performance of the methods on the testing dataset.

Method	AUC Coding	AUC Non-Coding	AUC All
PhD-SNP[g]	0.93	0.98	0.96
DANN	0.91	0.97	0.94
CADD	0.98	0.98	0.98
FATHMM	0.85	0.98	0.90
DeepSea	–	0.94	–
Eigen	0.93	0.82	0.96

AUC: Area Under the Receiving Operating Characteristic Curve.

Supplementary Figures

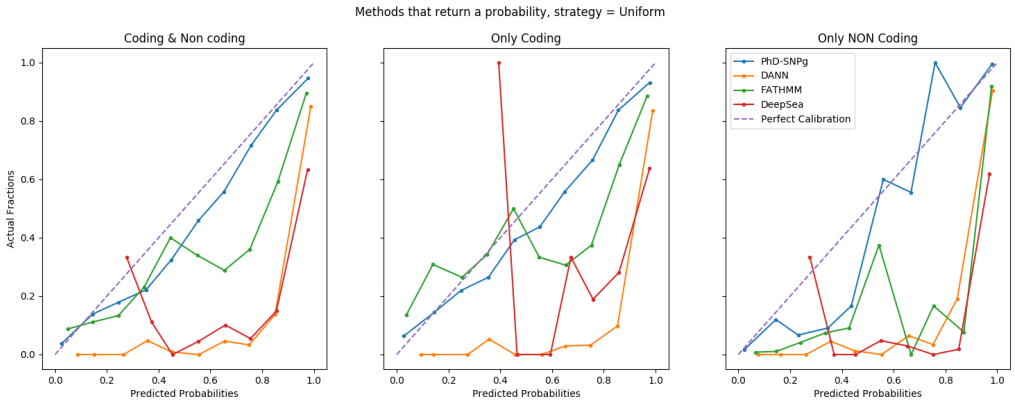

Figure S1. Calibration curves of the predictors that output a probability using the "uniform" strategy to plot the curves.

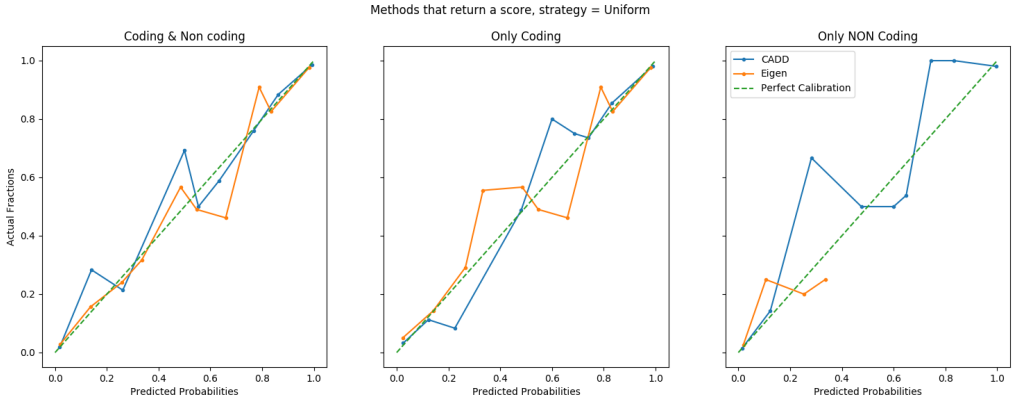

Figure S2. Calibration curves of the predictors that output a raw score using the "uniform" strategy to plot the curves. The scores were calibrated with the isotonic technique.

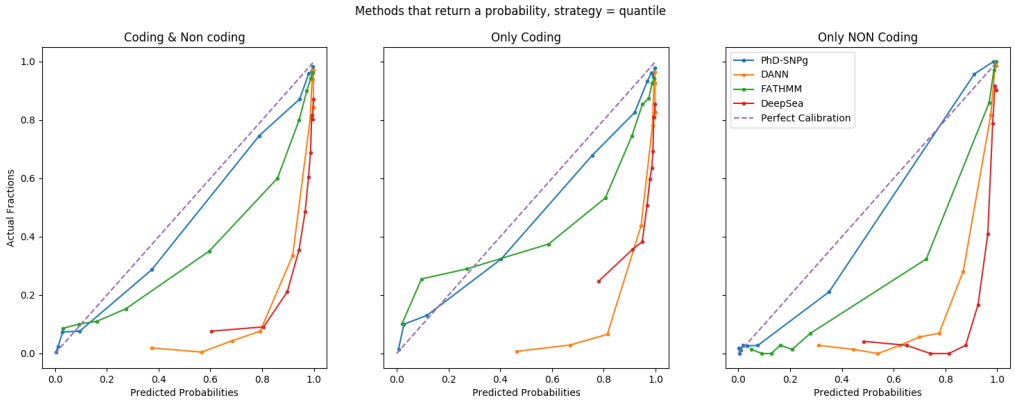

Figure S3. Calibration curves of the predictors that output a probability using the "quantile" strategy to plot the curves.

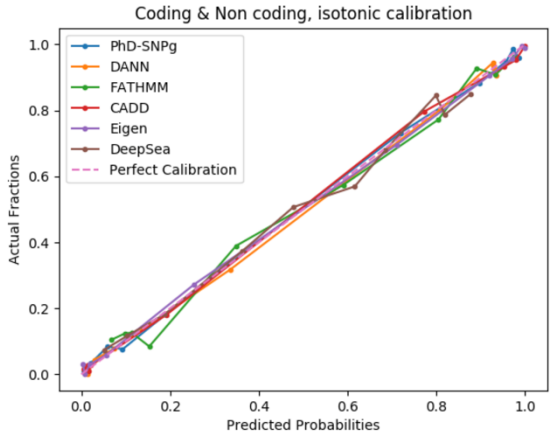

Fig. S4. Calibration curves of the predictors on coding and non-coding variants after isotonic calibration of their outputs.

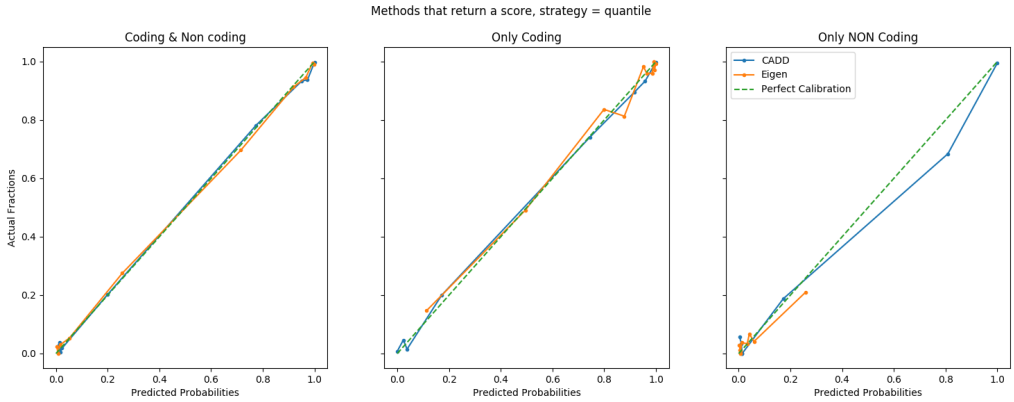

Figure S5. Calibration curves of the predictors that output a raw score using the "quantile" strategy to plot the curves. The scores were calibrated with an isotonic calibration.

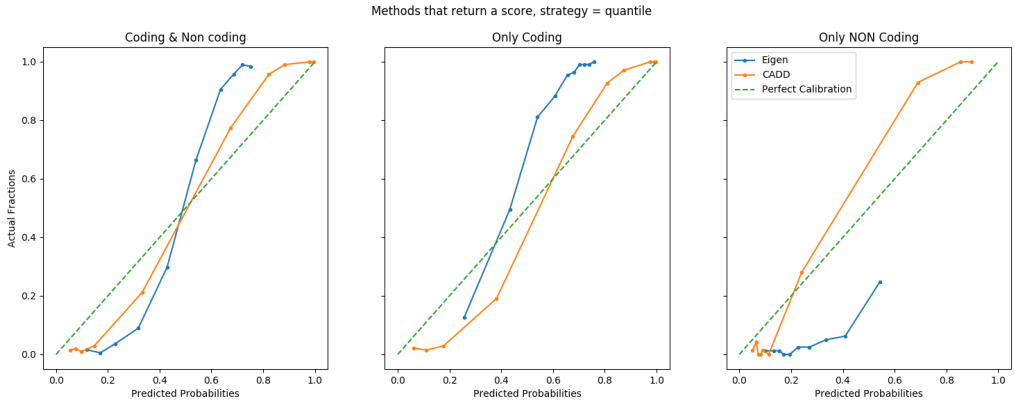

Figure S6. Calibration curves of the predictors that output a raw score using the "quantile" strategy to plot the curves. The scores were calibrated with a sigmoid calibration.

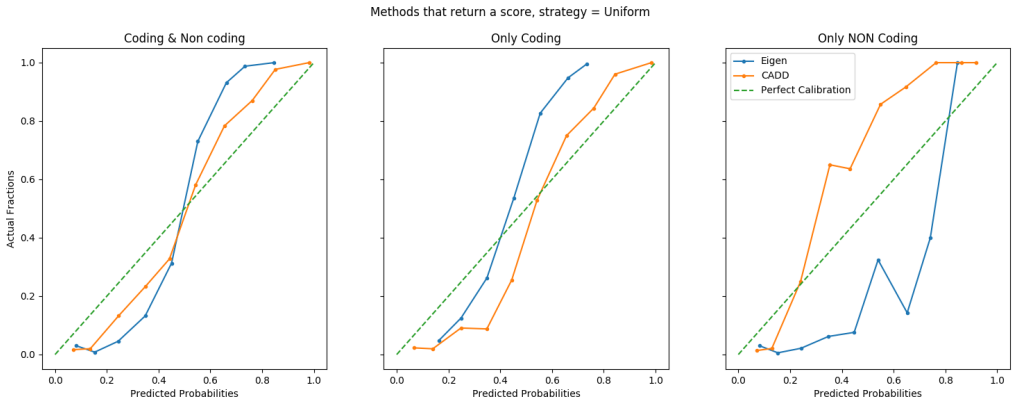

Figure S7. Calibration curves of the predictors that output a raw score using the "uniform" strategy to plot the curves. The scores were calibrated with a sigmoid calibration ($\frac{1}{1+e^{-Ax+B}}$). The best parameters were: A=1, B=2.5 for CADD and A=1, B=1.63/0.05 for Eigen.

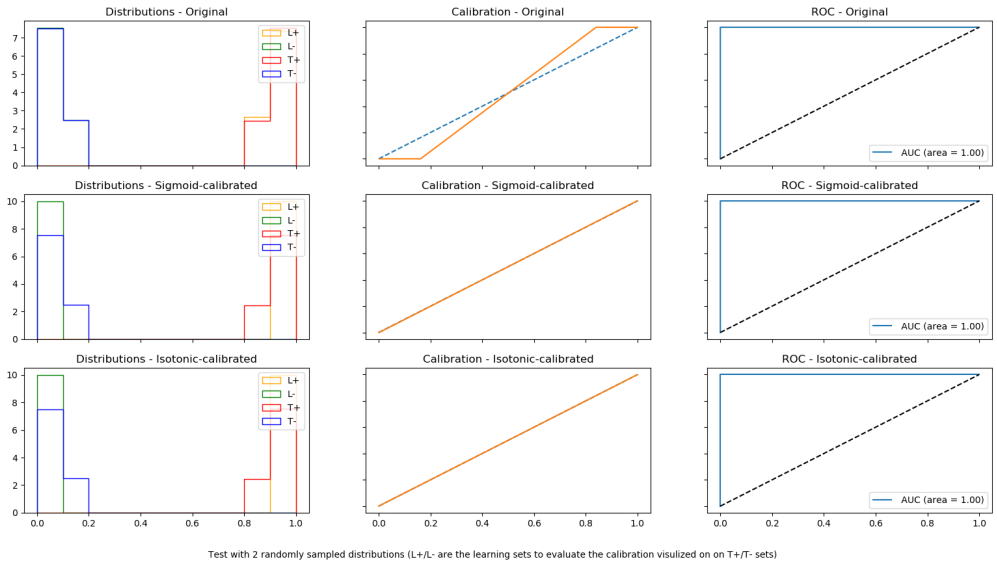

Figure S8. Simulated data: Perfect classifier, easy to calibrate both with isotonic and sigmoid techniques.

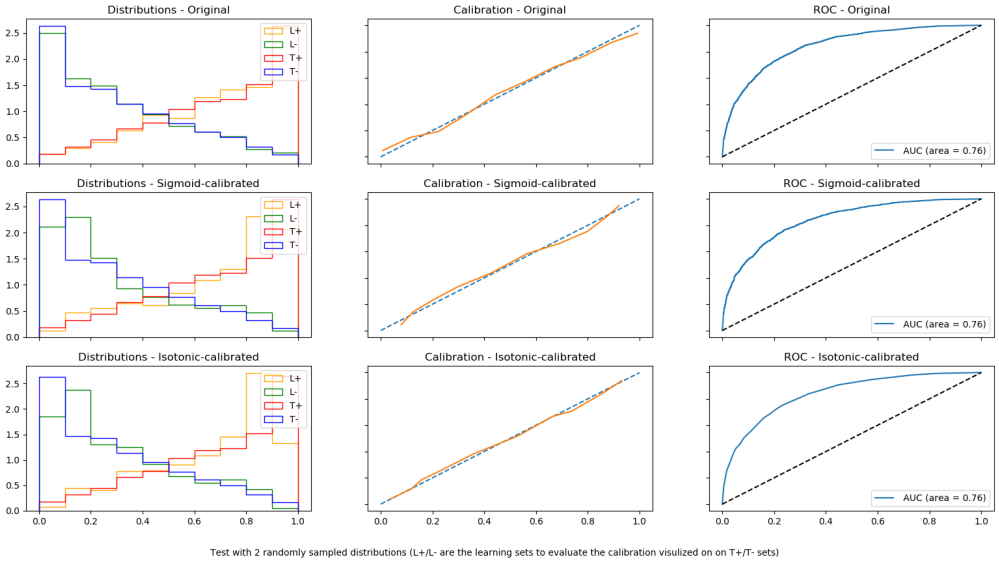

Figure S9. Simulated data: here is an example of a perfectly calibrated method that is not a perfect classier.

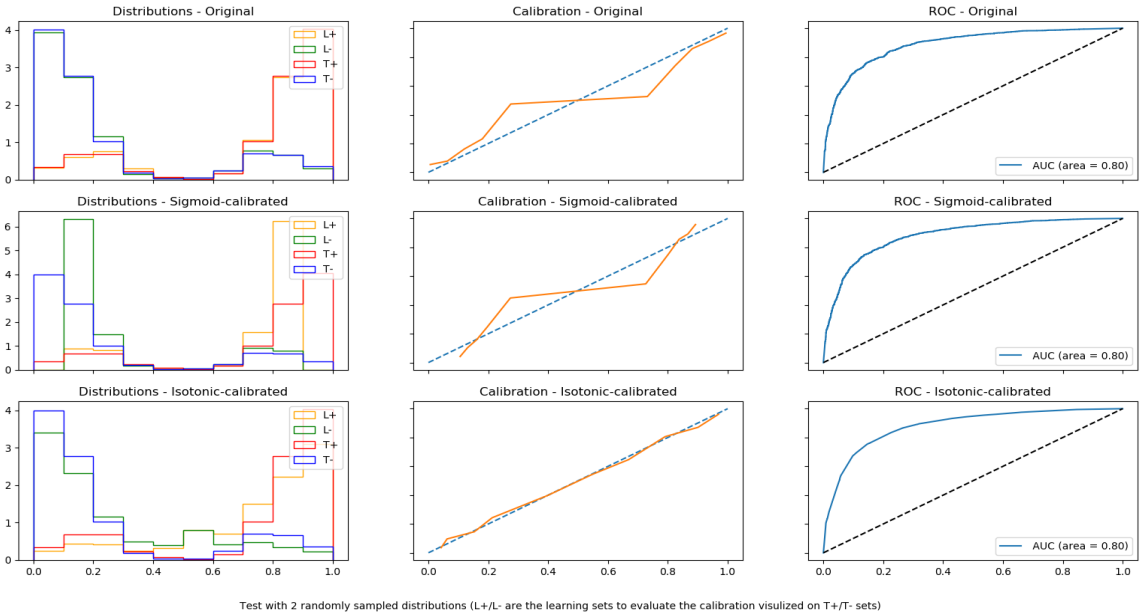

Figure S10. Simulated data: a harder situation to calibrate, that can be solved with the isotonic calibration.

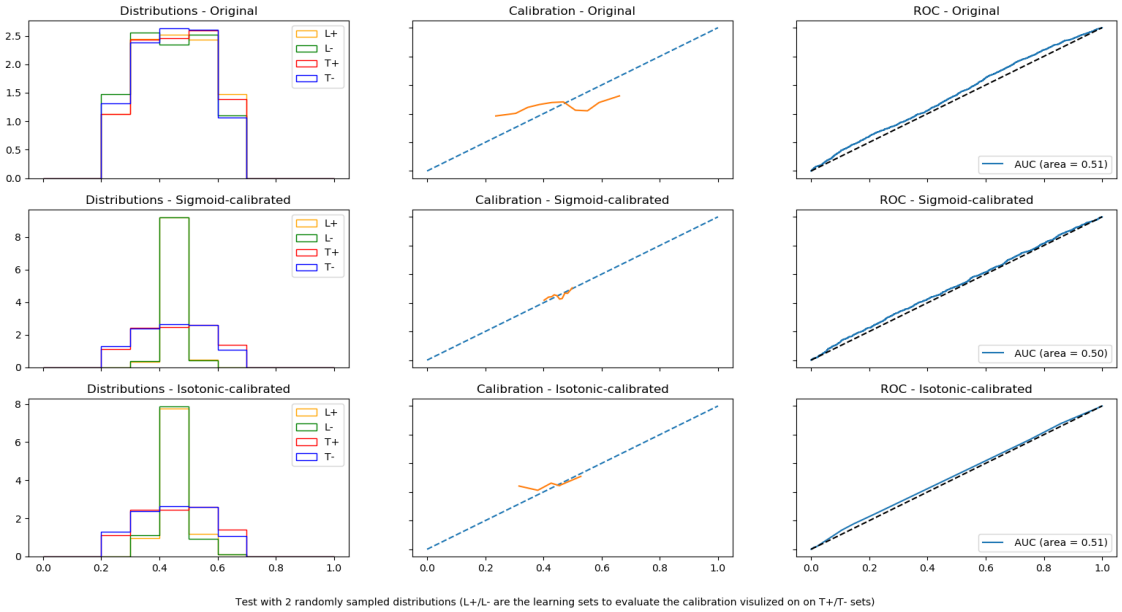

Figure S11. Simulated data: Random Classier.

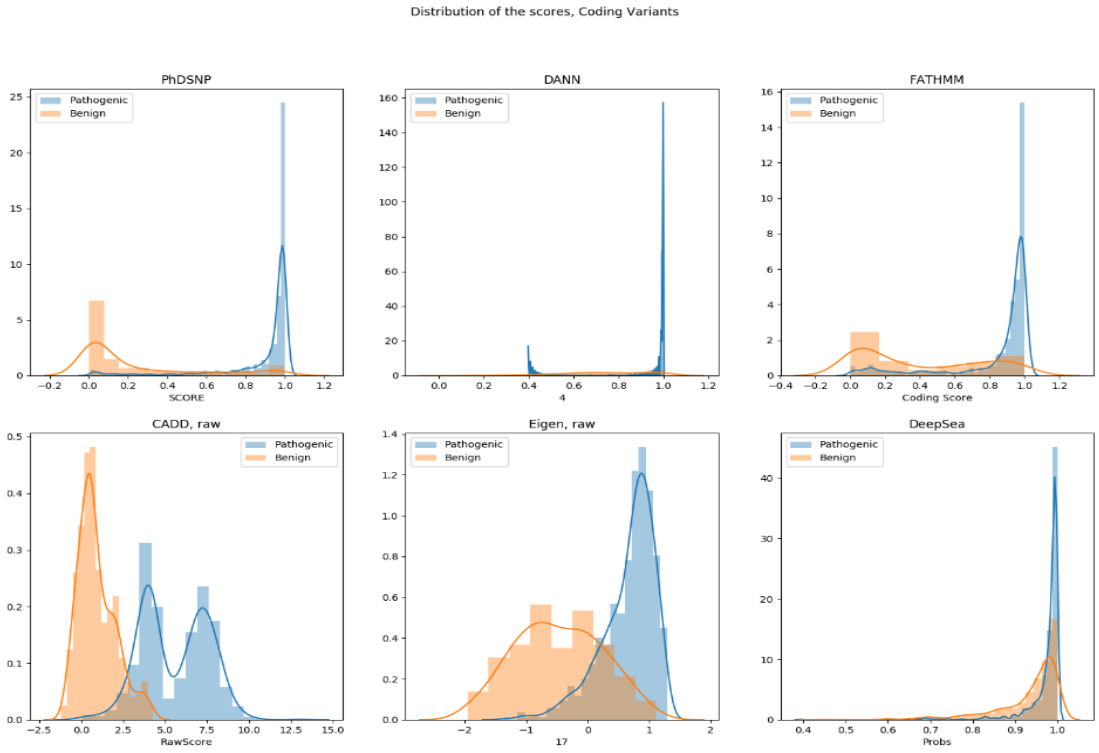

Figure S12. Distributions of the scores of the coding variants.

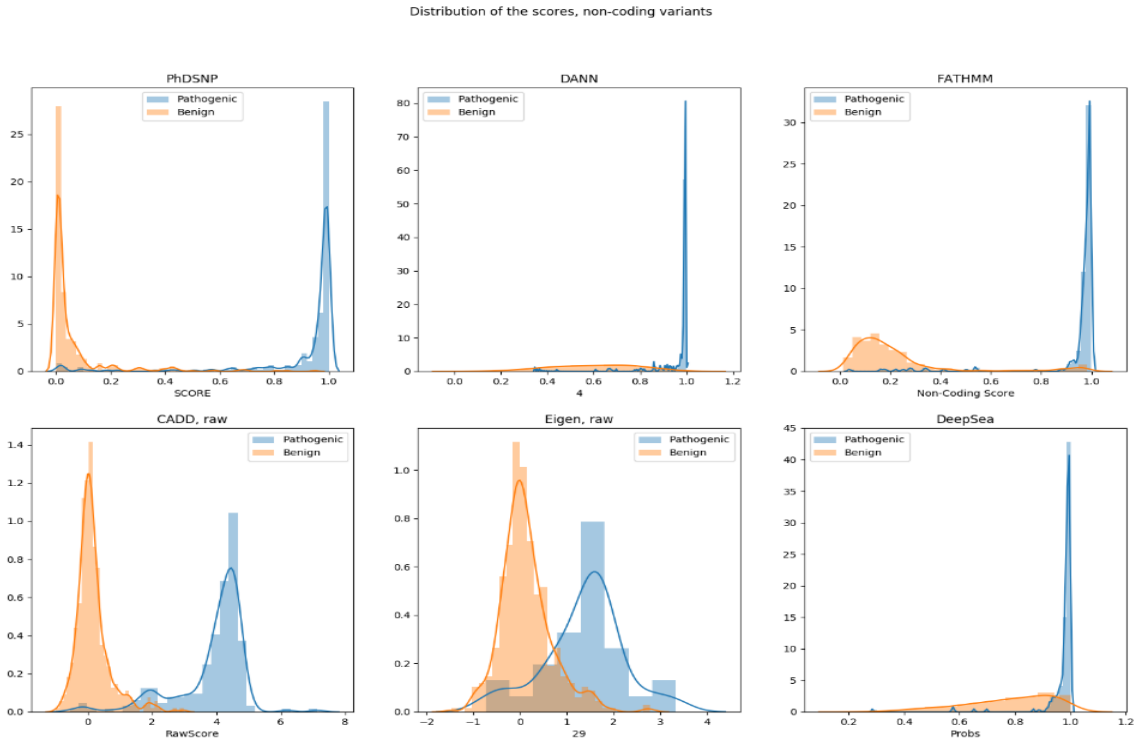

Figure S13. Distribution of the scores of the non-coding variants.

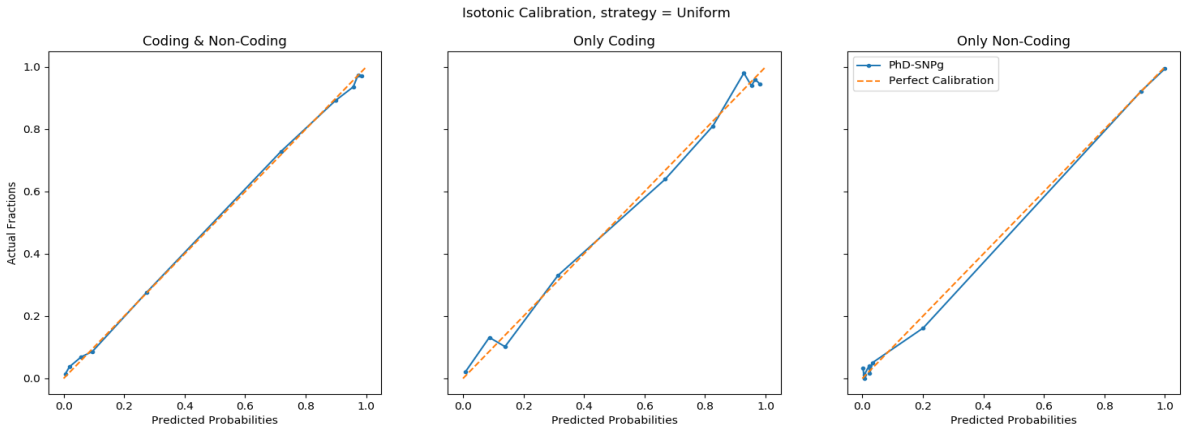

Figure S14. Calibration curves of PhD-SNP^{g} after isotonic calibration.

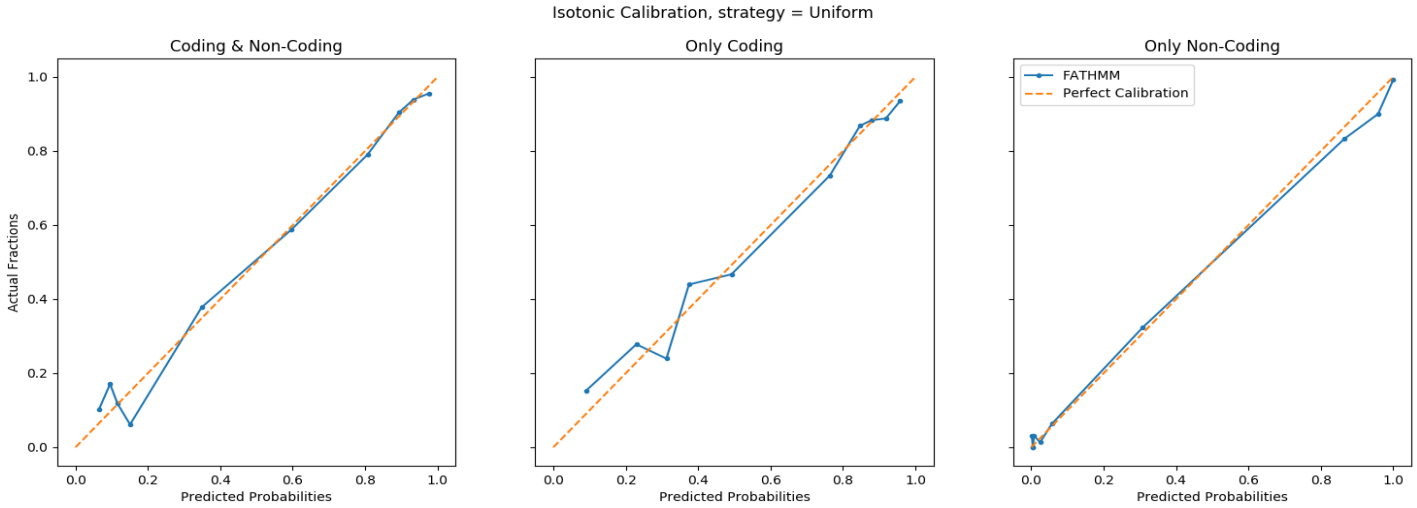

Figure S15. Calibration curves of FATHMM after isotonic calibration.

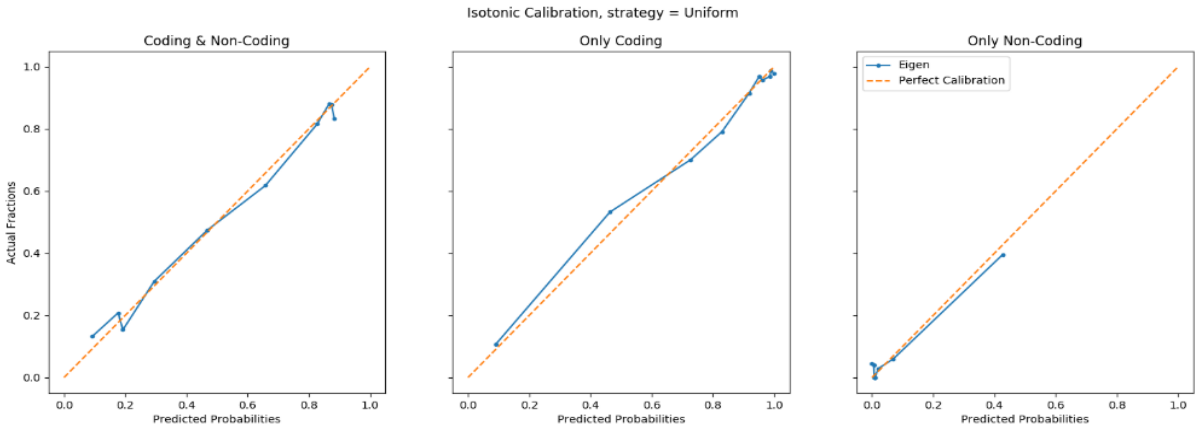

Figure S16. Calibration curves of Eigen after isotonic calibration.

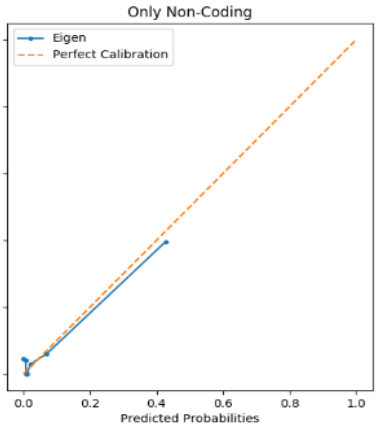

Figure S17. Calibration curve of DeepSea for non-coding variants after isotonic calibration.

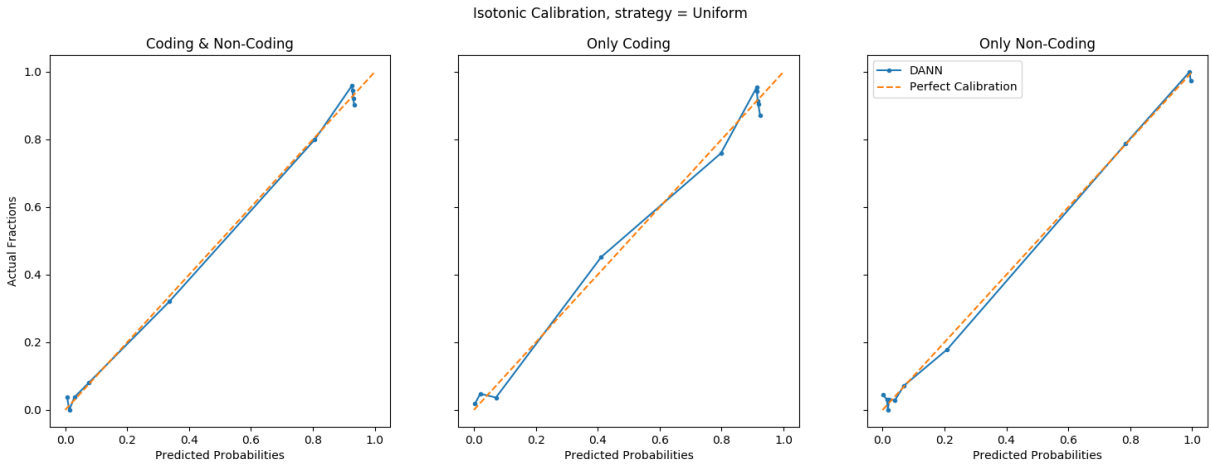

Figure S18. Calibration curves of DANN after isotonic calibration.

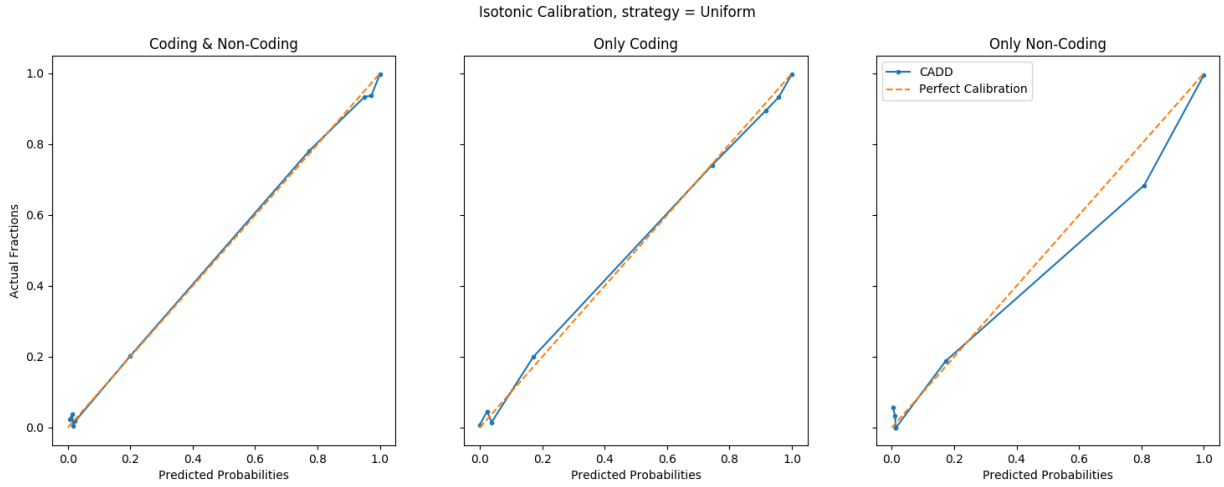

Figure S19. Calibration curves of CADD after isotonic calibration.

References

1000 Genomes Project Consortium*, et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;491(7422):56-65.

Brier, G.W. Verification of Forecasts Expressed in Terms of Probability,. *Monthly Weather Review* 1950;78:1-3.

Capriotti, E. and Fariselli, P. PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Res* 2017;45(W1):W247-W252.

Ionita-Laza, I.*, et al.* A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 2016;48(2):214-220.

Kircher, M.*, et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;46(3):310-315.

Landrum, M.J.*, et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016;44(D1):D862-868.

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In, *Proceedings of the 22nd International Conference on Machine Learning.*; 2005. p. 625-632.

Pedregosa, F. *et al.* (2011) Scikit-learn: Machine Learning in Python. *JMLR*, **12**, 2825–2830.

Quang, D., Chen, Y. and Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015;31(5):761-763.

Shihab, H.A.*, et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 2015;31(10):1536-1543.

Stenson, P.D.*, et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 2003;21(6):577-581.

Zhou, J. and Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;12(10):931-934.